Recall bias did not affect perceived magnitude of change in health-related functional status

Middel, B; Goudriaan, H; De Greef, M; Stewart, R; van Sonderen, E; Bouma, J; de Jongste, M

Published in:
Journal of Clinical Epidemiology

DOI:
10.1016/j.jclinepi.2005.08.018

Publication date:
2006

ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 503–511

Journal of Clinical Epidemiology

Recall bias did not affect perceived magnitude of change in health-related functional status

Berrie Middel[a,*], Heike Goudriaan[b], Mathieu de Greef[c], Roy Stewart[a], Eric van Sonderen[d], J. Bouma[d], Mike de Jongste[e]

[a]*Department of Health Sciences, University Medical Center Groningen, University of Groningen, P.O. Box 196, 9700 AD Groningen, The Netherlands*
[b]*Netherlands Institute for the Study of Crime and Law Enforcement, Leiden, The Netherlands*
[c]*Institute of Human Movement Sciences, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*
[d]*Northern Center for Healthcare Research, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*
[e]*Department of Cardiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*

Accepted 30 August 2005

Abstract

Background and Objective: It was hypothesized that within an invasively treated group and within a group that improved in angina pectoris no difference in effect size would occur between prospective and retrospective measures. Furthermore, it was hypothesized that assessment of perceived change at post-test may be invalid because of recall bias and present-state bias.

Study Design and Setting: Effect sizes (as standardized response means) were used as indicators of magnitude of change. Linear structural equation analysis (with LISREL) was used to investigate the relationship between the estimates of recall accuracy and retrospectively assessed change.

Results: No significant differences were found between prospective and retrospective measures of change over time in health-related functional status. Recall bias was not associated with retrospective measurement of change within a 12-week interval. An expected present-state effect was found in a structural equation model.

Conclusion: Prospective and retrospective indices of magnitude of change were similar between groups receiving treatment of known efficacy. Recall bias seems to be an acceptable risk in short-term follow-up studies.

Keywords: Health status indicators; Responsiveness; Prospective change; Heart failure; Retrospective change; Clinically relevant change; Effect size; Recall bias; Present-state bias

1. Introduction

The measurement of treatment-related change over a period of time in patients is central to both clinical research and practice. In evaluation research, investigators commonly define change as the difference between baseline and post-treatment scores, obtained from serial measurements (i.e., serial change). In clinical practice, however, change after treatment is generally assessed retrospectively by asking the patient to give an appraisal of the magnitude and direction of the change in health status or functioning as stable, improved, or deteriorated. In the interaction between clinician and patient, such a retrospective appraisal by the patient and physician concerning several domains of the health status has clinical relevance, in that it determines the decisions made in the management of the disease. This is common practice, and therefore consideration must be given to the possibility of measuring the retrospective change directly in evaluation studies of treatment efficacy with, for example, health-related functional status (HRFS) as the outcome.

In evaluation research, retrospective measurement is obviously easier and more economical than serial measurement. Despite the apparent advantages of retrospective measurement of change in HRFS, there is a suspicion that global or transition questions are biased due to recall problems or present-state effects at follow-up. It is assumed that prospective or serial change assessed by repeated measurement is superior and that the use of retrospective assessment of change in HFRS with global or transition questions is definitely not advisable [1].

* Corresponding author. Tel.: +31-50-3636504; fax: +31-50-3633059.
E-mail address: b.middel@med.rug.nl (B. Middel).

0895-4356/06/$ – see front matter
doi: 10.1016/j.jclinepi.2005.08.018

There is, however, an ongoing debate about the methods for estimating clinically relevant change [2–5]. One of the assumptions in this debate is that changes inferred from repeated measurement approximate the change captured by the patient's retrospective perceptions of change over a period of time [6–8]. Other researchers, however, have found that the retrospective recall of a change in health status or symptoms is not as accurate as the change found in pre–post designs because of the complexity of the question. For example, when an interviewer asks patients who have undergone a coronary artery bypass grafting (CABG) operation whether they have felt better or worse since the bypass operation, the patients have first to make a judgment of the present state of health, then make a reconstruction of the situation before the CABG, and finally do mental subtraction to estimate the perceived direction and amount of change over time.

This method has some weaknesses. First, there is often a correlation between the 'present state' score, the post-treatment score, and the 'retrospectively perceived change' score in that health status domain [9], because this post-treatment present state is the frame of reference for the comparison with the health status before the treatment—this is present-state bias. A second weakness is that when the time span is too long, people have great difficulty remembering how they were before treatment—this is recall bias. A third problem is that retrospective assessment of treatment-related change may be invalid if patients feel that they are being prevented from living as they would like to by problems not related to the disease for which they are being treated.

The fourth weakness is that patients who remained stable after an invasive operation (e.g., CABG), according to the outcome of repeated measurement, were obviously in some respects limited in functional status before this treatment, and consequently when posed a retrospective global question were likely to report improvement. Some transition items are too general (e.g., ''Have you felt better or worse since your bypass-operation?''). The patient may then refer only to a few symptoms manifesting themselves at that particular point in time, such as shortness of breath, pain in the chest, or fatigue [10–13]. Additionally, the single item is a relatively coarse method in comparison with the multi-item scale and is not as suitable for detecting the minor differences in health perception that may still be clinically relevant.

In the present study, multiple-item transition scales enable patients to rate the extent to which they have changed regarding a number of disease-specific variables, thereby allowing for the possibility that not all aspects of functioning, health status and symptoms will be given the same response. A scale constructed from the summed composite of transition items (transition scales) that belong to a HRFS domain—for example, physical, emotional, or social functioning—yields more information reflecting meaningful change in that dimension than single items would. Furthermore, the comparison of these retrospective change scales with multiple serial change scales, comprising identical items in terms of responsiveness, may contribute to the analysis of the convergent validity of prospective and retrospective measures of change in HRFS. In the present study, the importance or value that patients assign to their perceived change after treatment was used to weight the change in HRFS-scores-weighted items.

In an earlier publication based on the data from the present study [8], we showed that the serial change scores of items, and likewise the identical transition items of the physical functioning scale, yielded similar factor loadings and estimated internal consistency coefficients (Cronbach's α), despite the weaknesses of retrospective global questions. The results published by Aseltine et al. [14] correlate with our findings that no significant differences in responsiveness (standardized response mean: SRM) were observed between serial change scales and transition scales [8,15]. We therefore hypothesize that retrospective assessment of treatment-related change over time may not be affected by recall bias in short-term evaluation of medical treatment or interventions of approximately 6 weeks before and 12 weeks after a significant event or intervention such as percutaneous transluminal coronary angioplasty (PTCA) or CABG, and may be a valid and reliable proxy for serial change assessment with an HRFS measure.

The present study, however, explores the relationship between serial measurement and weighted or unweighted retrospective measurement, with identical items and scales belonging to the physical and emotional domains of health-related functional status. The patients in the present study were undergoing treatment that is known to have an impact on health status domains, and therefore the indices used should reflect meaningful change between baseline and follow-up.

The following questions were addressed in the present study:

1. Are responsiveness indices derived from serial change scores and from weighted and unweighted retrospective scores similar when patients are broken down into groups with known treatment efficacy?
2. To what extent is retrospective measurement using global questions influenced by recall bias?

2. Methods

2.1. Patient selection

To ensure that a change in health status manifested itself, we selected a group of patients undergoing treatment with a known efficacy and selected a disease-specific instrument with a known sensitivity to detect change over time [16]. The instrument had proved to be sensitive to change in a similar sample of Dutch patients with ischemic heart disease [17].

Patients participating in the present study were recruited consecutively from three hospitals in the north of the Netherlands. All the subjects were patients who, following a coronary angiography, were scheduled for PTCA or CABG, or who needed no operative intervention but received medication. Patients with other incapacitating diseases, cognitive impairments, aged 75 years or older, or who did not speak Dutch were excluded from the study. Ethical approval was obtained from the ethics committee at each participating hospital.

All participating patients received a mailed questionnaire accompanied by a written informed consent form. The questionnaire was prospectively administered at the baseline and 12 weeks after the decision for noninvasive intervention, or 12 weeks after the day of the PTCA/CABG intervention. We presumed that at baseline (and thus prior to the coronary angiography) both patients and cardiologists had no information about the subsequent decision concerning either intervention, and that this would therefore not affect the health status assessment and should reduce the risk of floor and ceiling effects. This control for potential bias resulted in logistical problems, however, and 6 months after the start of the study we were forced to select patients already waiting for outpatient treatment (PTCA) or waiting for hospital admission (CABG), shortly after the decision had already been taken by the cardiologists. The average time between baseline assessment and follow-up after CABG and PTCA was 15 weeks.

2.2. Measures

The Minnesota Living with Heart Failure Questionnaire (MLHF-Q) [18] is a disease-specific instrument that originally comprised 21 items. Two scales measure the physical functioning dimension (8 items), the emotional functioning dimension (5 items) and the overall score reflects health-related quality of life (21 items). In a previous Dutch sample [17], one item from the MLHF-Q had no correlation with the physical functioning factor as predefined by Rector and Cohn [18]. This also occurred in the present study [8], and so the item was not used in further analysis. Consequently, in the present study only the seven items for physical functioning and the five items for emotional functioning scales from the MLHF-Q were used. Eight separate MLHF-Q items were omitted in this methodological evaluation of serially and retrospectively assessed change in HRFS domains, because they consist of a heterogeneous set of social, financial, medical, and economic limitations. To investigate the concordance between serial change and perceived extent of change in HRFS domains, we extended the questionnaire with three items from the MOS-20 [19] physical functioning scale. These results have been reported elsewhere [5,8]. The response options were analogous to the MLHF-Q questionnaire's format (Fig. 1).

The physical dimension subscale has a range from 0 to 35 and the emotional dimension subscale from 0 to 25. The total score of the MOS-20 items assessing physical functioning ranges from 0 to 15. Consequently, the total scores of the physical functioning scale used in the present study range between 0 and 50. A higher score indicates a high negative impact of heart disease in the assessed aspects.

Scores of serial change items (SCI) were calculated by subtracting the follow-up score from the baseline score. A serial change score of zero was considered to indicate neither improvement nor deterioration and a high serial change score was indicative of a high degree of improvement.

To assess change in the same health domain using a straightforward method, we modified each item from the repeatedly assessed baseline questionnaire into direct questions of perceived change using global or transition questions [20]. Patients separately rated 15 items from the MLHF-Q and MOS-20 for perceived magnitude of change (PMC) and for the importance of that particular change in physical and emotional functioning in relation to their treatment.

The perceived magnitude of change questionnaire (Fig. 1) was used to classify patients according to whether they had improved, deteriorated, or perceived no change regarding each item of the questionnaire belonging to the dimensions of physical and emotional functioning. Patients who had undergone CABG or PTCA were asked, ''Since my operation, my problems with walking or climbing stairs in relation to my heart failure have become.…''; patients who had received medication were asked, ''Since the last time I filled out the questionnaire, my problems with…have become.…'' (and so on).

PMC was assessed after treatment at follow-up with these response categories: (1) increased greatly, (2) moderately increased, (3) increased a little, (4) has not changed, (5) decreased a little, (6) decreased moderately, and (7) decreased greatly. High scores in the transition scales indicate a high degree of improvement.

The estimated importance (I) of the perceived magnitude of change was elicited for each item with a 10-cm visual analog scale with anchors from 'least important' (rating = 1) to 'extremely important' (rating = 10). The individual patient ratings of perceived magnitude of change (PMC_i) and individual importance (I_i) for each HRFS item were multiplied together to form a weighted magnitude-importance score for each item ($WPMC_i$).

To give meaning to 'no change' in the calculation of the $WPMC_i$, the PMC_i item's original seven-point scores were expressed as deviations from 'no change' (score 4) by subtracting four points from each score : -3 = deteriorated greatly, -2 = deteriorated moderately, -1 = deteriorated a little, 0 = no change, 1 = improved a little, 2 = improved moderately, and 3 = improved greatly.

The total score for the patient's weighted magnitude of change-importance item scores was calculated by $\Sigma(PMC_i \times I_i)$. For each patient, a maximum possible score was calculated as the sum of the importance ratings (I_i) multiplied

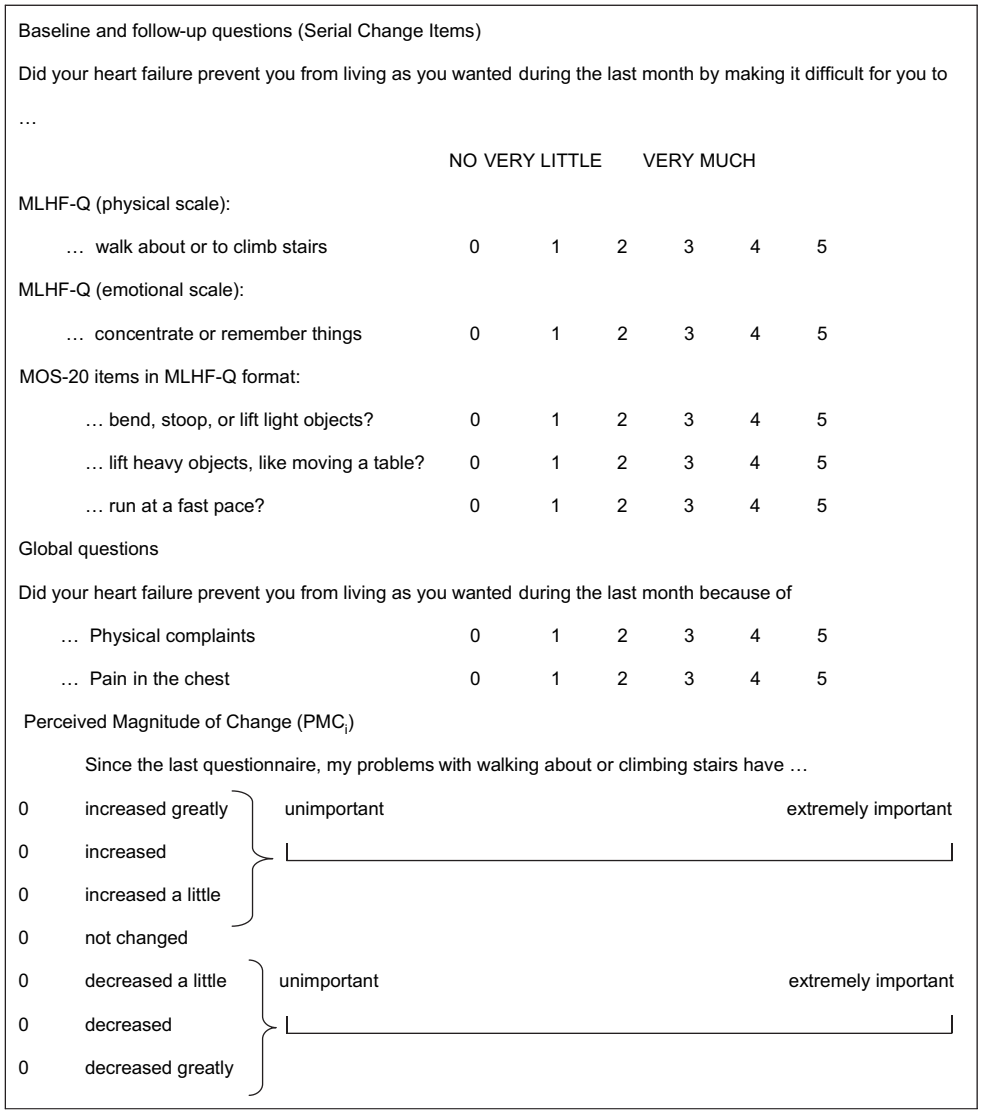

Fig. 1. Examples of questionnaire items.

by the maximum perceived magnitude of change score (PMC = 3).

The final patient-specific index (PSI) for each patient was estimated, following Wright and Young [21], as the ratio of the total sum of magnitude—that is, the importance scores divided by the maximum possible score times 100:

$$\mathrm{PSI} = \frac{\sum(\mathrm{PMC_i} \times I_i)}{\sum(3 \times I_i)} \times 100$$

To assess recall bias, the follow-up questionnaire contained two items asking patients to recall the extent to which they felt limited by physical problems or by pain due to their heart failure. The questions were introduced as follows: "If you were to assess your limitations due to pain before the operation (CABG/PTCA group), or since the first questionnaire you filled out some 12 weeks ago in the medication group, what would be your estimation?" The items were phrased as follows: "Before my operation, or at the first time I filled out the questionnaire, I was prevented from living as I wanted to because of the pain in my chest" or "…because of physical complaints." The response options ranged from 'no' (score = 0) and 'very little' (score = 1) to 'greatly' (score = 5) and were presented in the same format as the corresponding global question at baseline, phrased in the present tense (Fig. 1).

Subtraction of the item's score from the patient's score at baseline was used as an estimate of recall bias. The extent of deviation ranged from –5 to 5: a zero score indicates no recall bias and a score of absolute value 5 (i.e., either negative or positive) indicates the maximum amount of recall bias. Such a score of 5 indicates that the estimates made by the patient at 12-week follow-up regarding the extent of limitation before treatment, and the estimate actually given at baseline, were completely the opposite of each other.

2.3. Statistical methods

To make comparisons between the responsiveness of serial change scales (SCS), unweighted PMC items (UPMC), weighted WPMC items, and the summed PSI, responsiveness was quantified by means of SRM. The effect size of the serial measurement was calculated as the mean change in the measurements for the group divided by the standard deviation (SD) of the change scores of all the patients: SRM = Δ/SDΔ. For weighted and unweighted measures of perceived magnitude of change, responsiveness was considered to be the mean difference between no change divided by the SD of the difference, as previously reported by Fischer et al. [1]. Because 'no change' was scored as a zero for all the questions, it was simply calculated as the mean score measure for every group of patients divided by the SD of the measure for all patients. A higher SRM for a measure in a group with improved health-related quality of life indicates a greater ability of the instrument to detect a shift from the baseline.

A path model was analyzed to test the estimates of the magnitude of the effects of present-state bias and recall bias on retrospectively perceived change, and to estimate whether our data fit the proposed model. The analysis was performed with structural equation modeling using the maximum likelihood method (LISREL version 8.54) [22]. Residual correlations between physical health at baseline and follow-up and between serial change and recall bias were allowed, because standardized residuals indicated this correlation to exist. To allow for mutual comparisons between the path coefficients, the completely standardized solution was used. The fit of the model was evaluated by means of the comparative fit index (CFI) and the standardized root mean square residual (SRMR), in addition to the chi-square (χ^2) test. An adequate fit of the model is indicated by CFI $\geq$ .95, and SRMR $<$ .08 [23].

3. Results

3.1. Sample

A total of 398 candidates were screened for inclusion in the present study; 139 (35%) did not return the first questionnaire, so the final sample consisted of 259 patients. The probability of systematic differences between nonresponders and the study sample could not be tested, because no information was available without the written informed consent of the patients who did not return the first questionnaire.

Forty-two patients (16%) dropped out of the study before the follow-up assessment because of the death of the patient ($n = 7$) or because the patient had no heart failure ($n = 9$), refused further participation ($n = 9$), was too ill at follow-up ($n = 3$), had moved ($n = 3$), or did not react at all ($n = 11$). To ensure that the patients who dropped out at follow-up did not deviate systematically from the study group, the characteristics of these patients at the time they returned the first questionnaire were compared with the baseline characteristics of those who completed the questionnaire at follow-up. Demographic characteristics of the two groups were similar, except that the study sample had a statistically significant higher level of education ($\chi^2 = 14.70$, $P < .05$). This comparison also showed that there were no significant differences in the mean scores in the baseline health-status scales. Analyses were based on the 217 subjects who filled in the questionnaires at baseline and post-test.

The mean age of the patients was 60.6 years (SD = 9.43), with a range from 25 to 75. The gender breakdown was 61 female and 156 male. More men than women had a partner (92% vs. 72%), a higher education (18% vs. 7%), and employment (36% vs. 7%). At follow-up, 29% had undergone a CABG, 33% had a PTCA, and 38% had received pharmacotherapy. These and other characteristics are presented in Table 1.

3.2. Known groups validity

The comparison of serial change scales (SCS) with perceived magnitude of change (PMC), weighted perceived magnitude of change (WPMC), and PSI showed no differences in responsiveness estimates in the present study.

It was hypothesized that if a distinction between invasive and noninvasive treatment and between improvement and stability in angina pectoris was made, differences in the

Table 1
Sociodemographic characteristics of patients receiving three different treatments after coronary angiography

Characteristic	no. (%)
Gender	
Male	156 (72)
Female	61 (28)
Marital status	
Married	170 (78)
Cohabiting	15 (7)
Partner, not cohabiting	2 (1)
Unmarried	6 (3)
Divorced	8 (4)
Widow or widower	16 (7)
Employment status	
Employed	57 (26)
Unemployed or retired	147 (68)
Education	
Grade 6	44 (20)
Technical school (grades 7–9)	61 (28)
Junior high school (grades 7–9)	34 (16)
Junior high school including vocational education	33 (15)
High school or A levels	6 (3)
College (undergraduate study, 4 years)	22 (10)
University (graduate study, 5th year and beyond)	7 (3)
Treatment	
CABG	64 (29)
PTCA	71 (33)
Pharmacotherapy	82 (38)

Abbreviations: CABG, coronary artery bypass grafting; PTCA, percutaneous transluminal coronary angioplasty.

magnitude of the effect would be found between these groups. Invasive PTCA or CABG treatments were expected to produce more change, and noninvasive treatment was expected to produce very little change in HRFS over a period of time. From our responsiveness analysis (Table 2), the SRM of the physical function scale was .73 for SCS scores, .78 for PMC scales, and .75 for the WPMC scales. The PSI scores yielded the highest responsiveness index, .88. In the group of patients who improved in angina-related chest pain, effect sizes of the physical functioning scales were higher than in the invasively treated group, and in both groups an identical trend was reflected. The confidence intervals (95% CI) showed no overlap, indicating statistically significant different effect sizes between the invasively treated vs. noninvasively treated patients and significant different effect sizes between improved angina pectoris patients vs. stable patients. The SRMs of the SCS, PMC, and WPMC scales of emotional function in the invasively treated group ranged from .36 to .39 and the PSI emotional functioning scale yielded the highest effect size, .53.

The group of patients showing an improvement in angina-related chest pain showed larger effect sizes in the emotional functioning scales than the group treated with PTCA or CABG. The PSI again yielded the highest effect size. No differences were found between the SRMs from PMC and WPMC scales within each group compared to the SCS in both dimensions. As hypothesized, greater magnitude of change was found in the invasive treatment group. This was also the case in the group showing improvements in angina pectoris for SCS, PMC, WPMC, and PSI scales of physical and emotional functioning when compared to the noninvasively treated and stable groups, respectively. Furthermore, the SRM indices of the emotional functioning scale showed similar results, although smaller in magnitude.

3.3. Recall bias and present-state bias of change in health reevaluation

In the path model (Fig. 2), the following latent variables were used in the analysis: (1) the perceived magnitude of a change (PMC) in physical state of health and in angina-related chest pain, (2) the baseline physical state of health and in angina-related chest pain, (3) the present state of physical health and angina-related chest pain at follow-up, and (4) the recall of the physical state of health and angina-related chest pain. The actual differences between responses to global questions at baseline (e.g., ''Did your heart failure prevent you from living as you wanted to due to pain in the chest?'') and the matching question at follow-up (''Before my operation or the first time I filled out the questionnaire, I was prevented from living as I wanted to because of the pain in my chest'') are given in Table 3.

Deviations from zero (which represents no recall bias) are an indication of the recall bias magnitude. Pain in the chest, as a single health-related event, seems to be more accurately recalled than physical problems related to more than one health-related event.

Within the structural model, serial change was estimated through the latent variables representing the subtraction of physical functioning and pain on the chest at baseline from these outcomes at follow-up (i.e., present state). Recall bias was estimated using the latent variables representing the recall of baseline physical health state and pain at follow-up minus baseline. Therefore, to estimate the differences between baseline and follow-up and between recall at follow-up and the patient's estimate of the extent of limitation in functioning and pain in the chest at baseline, the path coefficients were fixed at 1 and −1. A structural equation model (SEM) was made to analyze whether retrospective measurement of perceived magnitude of change in HRFS may be influenced by recall bias or present-state bias. Path coefficients were estimated using the maximum likelihood method and represent the magnitude of the relationship between an item and a latent factor or between latent factors. A probability (*P*-value) of .10 indicates that ''the model's covariance matrix is sufficiently close to the observed data covariance matrix for the remaining differences to be mere sample fluctuations'' and a root mean square error of approximation (RMSEA) of .04 indicates an adequate fit of the model [23].

Table 2
Responsiveness as indicated by SRM for four measures in groups of patients differing in treatment impact and treatment effect

	Treatment		Angina-related chest pain		
	Invasive, n = 135	Noninvasive, n = 82	Improved, n = 87	Stable, n = 121	Overall, n = 217
Physical functioning scale, SRM (95% CI)					
SCS	.73 (.55, .91)	.29 (.09, .49)	.82 (.60, 1.04)	.35 (.17, .53)	.56 (.42, .70)
PMC	.78 (.60, .96)	.12 (−.06, .30)	.89 (.67, 1.11)	.28 (.12, .44)	.53 (.39, .67)
WPMC	.75 (.57, .93)	.10 (−.08, .28)	.86 (.64, 1.08)	.25 (.09, .41)	.51 (.37, .65)
PSI	.88 (.72, 1.04)	.30 (.03, .57)	1.00 (.80, 1.20)	.44 (.22, .66)	.70 (.56, .84)
Emotional functioning scale, SRM (95% CI)					
SCS	.39 (.21, .57)	.17 (−.05, .39)	.48 (.30, .66)	.14 (−.04, .32)	.31 (.17, .45)
PMC	.36 (.20, .52)	.18 (−.06, .42)	.51 (.29, .73)	.13 (−.05, .31)	.30 (.16, .44)
WPMC	.38 (.22, .54)	.19 (−.05, .43)	.52 (.30, .74)	.14 (−.04, .32)	.31 (.17, .45)
PSI	.53 (.33, .73)	.32 (.07, .57)	.76 (.54, .98)	.21 (−.01, .43)	.46 (.30, .62)

Abbreviations: CI, confidence interval; PMC, unweighted perceived magnitude of change scales; PSI, patient-specific index; SCS, serial change scales; SRM, standardized response mean; WPMC, weighted perceived magnitude of scales.

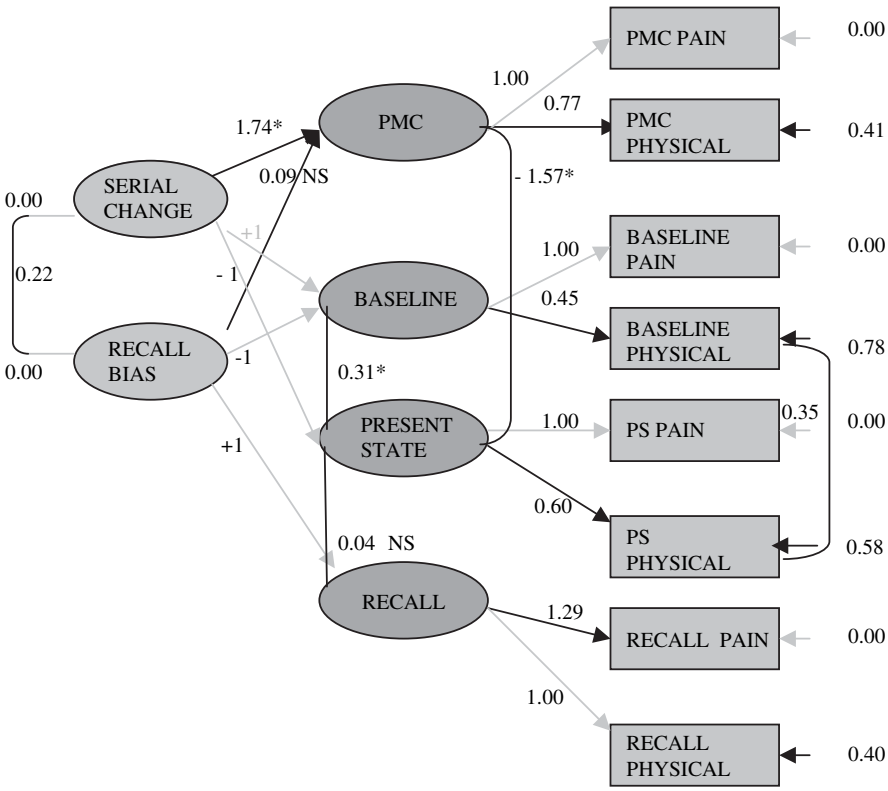

Fig. 2. LISREL path model of the relationship between recall bias and present-state bias and retrospectively perceived magnitude of change in physical problems and in angina-related pain of the chest. $\chi^2 = 24.84$, df = 17, $P = .10$, root mean square error of approximation RMSEA = .05 (*P <.05).

The model shows that the negative path coefficient (−1.57) between the magnitude of perceived change scale (PMC) after treatment, and the extent of limitation at follow-up (i.e., at the present state) is significantly larger than zero. Thus, higher scores on questions enquiring into the extent of limitation at follow-up due to angina-related chest pain and physical problems are associated with lower scores on perceived magnitude of change items in the same domains, indicating a deterioration. Only a nonsignificant relationship (.09) was found between the indicator of recall bias and the perceived magnitude of change. There was a significant relationship between recall of the physical problems, pain in the chest, and baseline (.31) and no significant relationship between these variables at the present state (.04). Earlier results [8] were confirmed by the finding of a significant relationship (1.74) between the indicators of serial change and perceived magnitude of change.

4. Discussion

There are many published studies that make use of items or global questions measuring the perceived magnitude of change in HRFS as measurements of treatment outcome [7,14,20,24]. In most of these studies, however, these questions are used as a single item to assess perceived change. One of the problems with globally assessed change is that the reliability cannot be established since Cronbach's α cannot be computed for a single item. The serial change scale (SCS), and the unweighted perceived magnitude of change (PMC) scale of physical functioning had a satisfactory level of internal consistency and yielded a Cronbach's α of .86 and .92, respectively [8].

The repeated measurement and retrospective methods applied in the present study have various strengths and weaknesses. The main problems with working with repeated measures and directly derived change scores are that they have a significant regression effect and are prone to

Table 3
Ranked recall bias scores on pain in the chest and physical problems

Recall bias score[a]	Pain in the chest (n = 214), %	Physical problems (n = 213), %
0	61.2	48.4
2	18.7	19.7
3	15.4	14.1
4	3.7	11.7
5	0.9	6.1
	Mean recall bias score (SD)	
	1.03 (1.38)	1.59 (1.71)

Abbreviation: SD, standard deviation.

[a] Recall bias score of 0 indicates identical score on either question (i.e., deviations from zero indicate recall bias). Scores above zero show the magnitude of the deviation on the six-point scale. For questions and scale, see Fig. 1. Recall bias score of 2 can thus be achieved by differing responses of 5 vs. 3, 4 vs. 2, 3 vs. 1, or 2 vs. 0. Similarly for 3 (5 vs. 2, 4 vs. 1, 3 vs. 0), 4 (5 vs. 1, 4 vs. 0), and 5 (5 vs. 0).

measurement error [25–27]. Change scores derived from repeated measurement may also be flawed by floor and ceiling effects [28,29]—for example, carryover effects of learning if the retest intervals are too short, or because of specific events occurring between the first and second measurement, the natural course of the disease, or acquiescence and social desirability [30].

Another threat to the validity of change scores is the assumption made by researchers that subjects have an internalized perception of their level of functioning with regard to, for example, the domain of physical health status, and that this internalized standard will not change between baseline and post-test. Respondents in the present study may also have recalibrated their baseline situation due to the clinical intervention. For example, because of their high hopes for a better physical state of health they may have felt inclined to give socially desirable answers to health professionals, or they may have changed the anchors for their ratings over time due to significant and substantial changes in the course of the disease. These confounding factors are associated with what is called response shift [29,31,32].

There are also several threats to the reliability and validity of our findings based on the use of multitransition items in the present study. Retrospective perception of change may not be accurate, due to recall bias, and patients may equate their present state with a change in their health status. A respondent doing poorly after treatment may be inclined to think that in general things are getting worse, even with improved or unchanged health state [9,14,24]. In the study by Fitzpatrick et al. [33], however, the transition questions were shown not to be determined by the patient's mood at post-test nor by the present state.

Patients in the present study who suffered a significant decline in health due to invasive treatment such as CABG of PTCA may also have overestimated their perception of baseline health if they were longing for the time when their health was better [14,29]. We decided to use a short time interval, because inaccurate recall seems to be determined by a lengthy time interval, exposure or intervention, and by the degree of detail required [34].

The significance, the vividness, and meaningfulness of events also contribute to a more accurate recall, and this was the case in the present study. Some of the findings published by Aseltine et al. [14] suggest that their measures assessing more concrete aspects of a patient's condition provided greater correspondence between prospective and retrospective assessment than the more abstract measures of general health. Despite the limitations of transition questions, there is a growing realization that patients can be more directly involved in judging for themselves whether treatments have improved their health status in relation to the observed health status of other patients by directly asked transition questions [28,33,35–37]. Moreover, transition questions were shown to be more sensitive to changes over time in health-related quality of life than change scores [1,2,6].

Central to our analysis was the assumption that the multi-item transition indices (scales) measure magnitude of change similar to that estimated with the serially assessed change in domains of health to which they were paired. In several studies, the agreement between retrospective assessments and serial assessments was poor if single items were used. The results of the present study therefore argue for multi-item batteries of transition items, because single-item transition questions do not cover a representative sample of the health aspects that belong to the underlying construct or dimension. Further studies should be conducted to address the psychometric aspects of transition scales used repeatedly in longitudinal studies, such as test–retest reliability or the ability to discriminate between known groups.

The retrospective judgment of change is difficult. Patients must be able to quantify both their present state and their state at baseline and then perform a mental subtraction. There is evidence that patients are in fact unable to remember their initial state, and that the judgment is based on an implicit theory of change beginning with their present state and working backwards [24]. This typically results in a high positive correlation between the retrospective scores and the present state, and a correlation near zero between the retrospective scores and the baseline measure.

The correlation between the present-state questions and concordant transition questions seems logical in a sample of patients who underwent treatment with a known efficacy. In the present study, it was expected that a perceived improvement in, for example, climbing stairs should correlate with no limitations in climbing stairs after PTCA or CABG, because these treatments aim to improve the physical condition of climbing stairs. Patients first completed the present-state questionnaire, which was followed by questions on retrospectively perceived change. To control for present-state bias, patients should be randomly assigned to repeated measurement or serial prospective measurement of change or to retrospective measurement with transition questions.

In the present study, it was assumed that the scores on outcome measures at baseline may have been affected in patients with knowledge of the treatment plan or who are scheduled on a waiting list for CABG or PTCA compared with patients who had no knowledge of what was going to happen. Due to logistical problems, only 67 patients were blinded with regard to the plan of treatment. It was expected that patient who were blinded would have lower psychological stress, in particular those who underwent an invasive treatment. Baseline scores on the emotional functioning scale of these 67 patients (mean = 69.7, SD = 25.2) did not differ significantly from the 250 patients who were not blinded (mean = 63.3, SD = 27.6), with a 95% CI from −13.7 to 0.94. We therefore do not expect that patients knowingly going for invasive treatment will experience a different change in emotional functioning from those who go blinded into it.

To our knowledge, no other studies have used a set of PMC or WPMC transition items to measure change in domains of health such as physical functioning or emotional functioning. Despite the shortcomings of the present study, we believe that it is important for further research. The use of sets of multiple-item transition scales to measure change in health domains provides an opportunity for an unequivocal representation of changes that are relevant for the patient. This method may also be considered in study designs where repeated measurement is not feasible, such as in the assessment of change in patients who have had an acute heart attack after emergency referral to a hospital.

References

[1] Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. JAMA 1999;282:1157–63.

[2] Ziebland S. Measuring changes in health status. In: Jenkinson C, editor. Measuring health and medical outcomes. London: UCL Press; 1994.

[3] Mahler DA, Weinberg DH, Wells CK, Feinstein AR. The measurement of dyspnea: contents, interobserver agreement, and physiologic correlates of two new clinical indexes. Chest 1984;85:751–8.

[4] Osoba D. Interpreting the meaningfulness of change in health-related quality of life scores: lessons from studies in adults. Int J Cancer Suppl 1999;12:132–7.

[5] Middel B, Stewart R, Bouma J, Van Sonderen E, van den Heuvel WJA. How to validate clinically important change in health-related functional status: Is the magnitude of the effect size consistently related to magnitude of change as indicated by a global question rating? J Eval Clin Pract 2001;7:399–410.

[6] Ziebland S, Fitzpatrick R, Jenkinson C, Mowat A, Mowat A. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the Health Assessment Questionnaire (HAQ) and modified HAQ. Ann Rheum Dis 1992;51:1202–5.

[7] Emery CF, Blumenthal JA. Perceived change among participants in an exercise program for older adults. Gerontologist 1990;30:516–21.

[8] Middel B, de Greef M, de Jongste MJL, Crijns HJGM, Stewart R, van den Heuvel WJA. Why don't we ask patients with coronary heart disease directly how much they have changed after treatment? J Cardiopulm Rehabil 2002;22:47–52.

[9] Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 2nd ed. Oxford: Oxford University Press; 1995.

[10] Kempen GIJM, Miedema I, van den Bos GAM, Ormel J. Relationship of domain-specific measures of health to perceived overall health among older subjects. J Clin Epidemiol 1998;51:11–8.

[11] Read JL, Quin RJ, Hoefer MA. Measuring overall health: an evaluation of three important approaches. J Chron Dis 1987;40(Suppl):7S–19S.

[12] Leavey R, Wilkin D. A comparison of two health survey measures of health status. Soc Sci Med 1988;27:269–75.

[13] Kempen GIJM. The MOS Short-Form General Health Survey: single item vs. multiple measures of health-related quality of life: some nuances. Psychol Rep 1992;70:608–10.

[14] Aseltine RH Jr, Carlson KJ, Fowler FJ Jr, Barry MJ. Comparing prospective and retrospective measures of treatment outcomes. Med Care 1995;33(4 Suppl):AS67–76.

[15] Middel B. Assessment of change in clinical evaluation. Groningen: Northern Centre for Healthcare Research, University of Groningen; 2001.

[16] Guyatt GH. Measurement of health-related quality of life in heart failure. J Am Coll Cardiol 1993;22(4 Suppl A):185A–91A.

[17] Middel B, Bouma J, de Jongste M, Van Sonderen E, Niemeijer MG, Crijns H, van den Heuvel W. Psychometric properties of the Minnesota Living with Heart Failure Questionnaire (MLHF-Q). Clin Rehabil 2001;15:489–500.

[18] Rector TS, Cohn JN; Pimobendan Multicenter Research Group. Assessment of patient outcome with the Minnesota Living with Heart Failure questionnaire: reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. Am Heart J 1992;124:1017–25.

[19] Stewart AL, Hays RD, Ware JE. The MOS Short-form General Health Survey: reliability and validity in a patient population. Med Care 1988;26:724–35.

[20] Perneger TV, Etter J-F, Rougemont A. Prospective versus retrospective measurement of change in health status: a community based study in Geneva, Switzerland. J Epidemiol Community Health 1997;51:320–5.

[21] Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239–46.

[22] Jöreskog KG, Sörbom D. LISREL-8: user's reference guide. 2nd ed. Chicago: Scientific Software International; 2003.

[23] Hu L, Bentler PM. Fit indices in covariance structure modelling: sensitivity to underparameterized model misspecifications. Psychol Methods 1998;3:424–53.

[24] Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50:869–79.

[25] Nunnally JC. The study of change in evaluation research: principles concerning measurement, experimental design, and analysis. In: Struening EL, Brewer MB, editors. Handbook of evaluation research. Beverly Hills, CA: SAGE Publications; 1983:231–69.

[26] Gottman JM, Rushe RH. The analysis of change: issues, fallacies, and new ideas. J Consult Clin Psychol 1993;61:907–10.

[27] Hsu LM. Regression toward the mean associated with measurement error and the identification of improvement and deterioration in psychotherapy. J Consult Clin Psychol 1995;63:141–4.

[28] Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients: the floor phenomenon. Med Care 1990;28:1142–52.

[29] Baker DW, Hays RD, Brook RH. Understanding changes in health status: is the floor phenomenon merely the last step of the staircase? Med Care 1997;35:1–15.

[30] Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Chicago: Rand McNally College Publishing; 1979.

[31] Mancuso CA, Charlson ME. Does recollection error threaten the validity of cross-sectional studies of effectiveness? Med Care 1995;33(4 Suppl):AS77–88.

[32] Schwartz CE, Sprangers MAG. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc Sci Med 1999;48:1531–48.

[33] Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Transition questions to assess outcome in rheumatoid arthritis. Br J Rheumatol 1993;32:807–11.

[34] Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol 1990;43:87–91.

[35] Fitzpatrick R, Albrecht G. The plausibility of quality-of-life measures in different domains of health care. In: Nordenfelt L, editor. Concepts and measurements of quality of life in health care. Dordrecht: Kluwer Academic Publishers; 1994:201–27.

[36] Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. J Clin Epidemiol 1996;49:1215–9.

[37] Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements: an illustration in rheumatology. Arch Intern Med 1993;153:1337–42.